PATENT ABSTRACTS OF JAPAN

(11)Publication number : **10-078964**
(43)Date of publication of application : **24.03.1998**

(51)Int.Cl. **G06F 17/27**
G10L 3/00

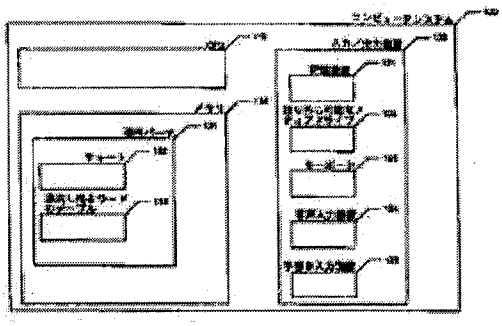
(21)Application number : **09-162383**
(22)Date of filing : **19.06.1997**

(71)Applicant : **MICROSOFT CORP**
(72)Inventor : **STEPHEN DAROU RICHARDSON**
GEORGE E HEIDOON

(30)Priority
Priority number : **96 671203** Priority date : **25.06.1996** Priority country : **US**

(54) METHOD AND SYSTEM FOR IDENTIFYING AND ANALYZING GENERALLY CONFUSED WORD BY NATURAL LANGUAGE PARSER

(57)Abstract:
PROBLEM TO BE SOLVED: To perfectly parse a sentence including the words confused with each other by using a list of confused word sets and processing the word encountered in an input sentence to show a speech part that is shown in another word of the same set as the encountered word.
SOLUTION: An input/output device 120 of a computer system 100 includes a keyboard 123 and also optionally includes a voice input device 124 and a handwritten input device 125 which are used by a user to indirectly input a natural language text. A parser 131 of a memory 130 identifies and analyzes the words which are generally confused with each other. The parser 131 contains a chart 132 including a parse tree which shows the input text segments and the intermediate parsing results and also contains a table 133 of the generally confused words and the words that can be confused and mapped to the generally confused words (i.e., the intentional words).



(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号

特開平10－78964

(43)公開日　平成10年(1998)３月24日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/27			Ｇ０６Ｆ　15/38	Ｄ	
Ｇ１０Ｌ　3/00	５６１		Ｇ１０Ｌ　3/00	５６１Ｇ	
			Ｇ０６Ｆ　15/20	５５０Ｊ	

審査請求　未請求　請求項の数10　ＯＬ　（全 15 頁）

(21)出願番号　特願平9－162383

(22)出願日　平成９年(1997)６月19日

(31)優先権主張番号　０８／６７１２０３

(32)優先日　1996年６月25日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　391055933
マイクロソフト　コーポレイション
ＭＩＣＲＯＳＯＦＴ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国　ワシントン州　98052－6399　レッドモンド　ワン　マイクロソフト　ウェイ　（番地なし）

(72)発明者　スティーヴン　ダロウ　リチャードソン
アメリカ合衆国　ワシントン州　98052　レッドモンド　ノースイースト　ワンハンドレッドアンドサーティセカンド　18028

(74)代理人　弁理士　中村　稔　（外７名）

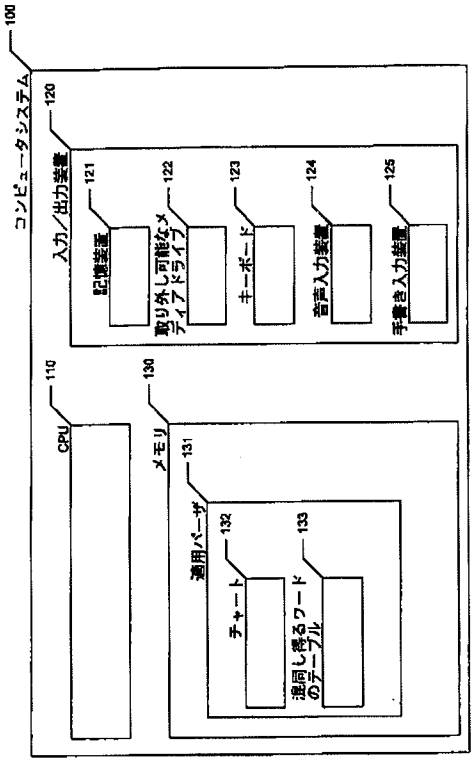
最終頁に続く

(54)【発明の名称】　一般に混同するワードを自然言語パーザにおいて識別及び分析する方法及びシステム

(57)【要約】

【課題】　自然言語パーザにおいて一般に混同するワードを識別しそして分析する方法及びシステムを提供する。

【解決手段】　コンピュータシステムは、２つ以上のワードより成る入力テキストを、入力テキストのワードの中の１つのワードを含む潜在的に混同するワードからその意図されたワードへとマップする関係を使用してパーズする。コンピュータシステムは、先ず、潜在的に混同するワードを含む入力テキストの各ワードに対して考えられるスピーチ部分を識別し、次いで、上記関係が潜在的に混同するワードをマップするところの意図されたワードに対して考えられるスピーチ部分を識別し、そしてそれらの識別されたスピーチ部分に構文的文法ルールを適用し、意図されたワードに対するスピーチ部分を含む完全な構文ツリーは発生されるが、潜在的の混同するワードに対するスピーチ部分を含む完全な構文ツリーは発生されないようにする。



1

【特許請求の範囲】

【請求項１】　コンピュータシステムにおいて１つ以上のワードを含む自然言語入力テキストのセグメントを文法ルールと複数のエントリーを含む辞書とを用いてパーズする方法であって、各辞書エントリーは自然言語のワードに対応しそしてそのワードに対する１つ以上の考えられるスピーチ部分を指定し、上記方法は、

（ａ）入力テキストセグメントを表すパーズツリーと、それに対する中間のパーズ結果とを含むチャートを形成し、

（ｂ）入力テキストセグメントに生じる各ワードごとに、そのワードの辞書エントリーで指定されたスピーチの部分を指定するスピーチ部分記録をそのワードに対して上記チャートに形成し、

（ｃ）入力テキストセグメントに生じるワードであって、別のワードと一般に混同するワードを識別し、

（ｄ）その識別されたワードと一般に混同するワードに対する辞書エントリーで指定されたスピーチの部分を指定するスピーチ部分記録を上記識別されたワードに対して上記チャートに形成し、そして

（ｅ）上記段階（ｂ）及び（ｄ）で形成された両方にスピーチ部分記録に文法ルールを適用する、という段階を備えたことを特徴とする方法。

【請求項２】　更に、一般に混同するワードのリストを使用し、このリストは、一般に混同するワードの各々に対し、そのワードと一般に混同するワードを含み、そして上記段階（ｃ）は、入力テキストセグメントに生じるワードの１つを上記リストのワードの１つとマッチングさせる段階を含む請求項１に記載の方法。

【請求項３】　上記段階（ｂ）は、入力テキストセグメントに生じる各ワードごとに、そのワードの辞書エントリーで指定された考えられるスピーチ部分の各々を指定するスピーチ部分記録を上記チャートに形成し、そして上記方法は、更に、入力テキストセグメントの各ワードごとに、そのワードに対してチャートに形成されたスピーチ部分記録を一緒にリンクする段階を備え、１つ以上の文法ルールをスピーチ部分記録に適用することは、そのスピーチ部分記録がリンクされる他のスピーチ部分記録を検査することによりそのワードに対して他の考えられるスピーチ部分を決定することを含み、そして更に、上記方法は、段階（ｄ）で形成されたスピーチ部分記録を段階（ｂ）で識別されたワードに対して形成されたスピーチ部分記録にリンクさせる段階を含む請求項１に記載の方法。

【請求項４】　上記段階（ｄ）は、上記段階（ｅ）の実行を開始した後に行う請求項１に記載の方法。

【請求項５】　上記段階（ｄ）は、上記段階（ｂ）で形成されたスピーチ部分記録への文法ルールの適用が終了した後に行う請求項１に記載の方法。

【請求項６】　各文法ルール及びスピーチ部分記録は、

2

適用優先値が関連され、上記段階（ｅ）は、適用優先値が減少する順に文法ルール及びリストのスピーチ部分記録を適用し、そして上記識別されたワードと一般に混同するワードに対するスピーチ部分記録に関連した適用優先値は、上記識別されたワードに対するスピーチ部分記録に関連した適用優先値より小さい請求項１に記載の方法。

【請求項７】　上記段階（ｅ）の文法ルールの適用が、上記識別されたワードを含む入力テキストの完全なパーズを形成しないが、一般に混同するワードを含む入力テキストの完全なパーズを形成するときには、上記識別されたワードが上記一般に混同するワードと混同することを指示する段階を更に備えた請求項１に記載の方法。

【請求項８】　上記段階（ｅ）の文法ルールの適用が、上記識別されたワードを含む入力テキストの完全なパーズを形成しないか、又は一般に混同するワードを含む入力テキストの完全なパーズを形成する場合には、

自然言語センテンスが構文的に正しくないという指示を出力し、そして自然言語センテンスの識別されたワードが、その識別されたワードと一般に混同するワードと置き換えられた場合には、自然言語センテンスが構文的に正しいという指示を出力する、という段階を備えた請求項１に記載の方法。

【請求項９】　１つ以上のワードを含む自然言語入力テキストのセグメントを文法ルールと複数のエントリーを含む辞書とを使用してパーズするための装置であって、各辞書エントリーは自然言語のワードに対応しそしてそのワードに対する１つ以上の考えられるスピーチ部分を指定し、上記装置は、

入力テキストセグメントを表すパーズツリーと、それに対する中間のパーズ結果とを含むデータ構造体と、

入力テキストセグメントに生じる各ワードごとに、そのワードの辞書エントリーで指定されたスピーチの部分を各々指定するスピーチ部分記録を上記データ構造体に形成する一次スピーチ部分記録発生器と、

入力テキストセグメントに生じるワードであって、別のワードと一般に混同するワードを識別する識別子と、

その識別されたワードと一般に混同するワードに対する辞書エントリーで指定されたスピーチの部分を指定するスピーチ部分記録を、上記識別子で識別されたワードに対してチャートメモリに形成する二次スピーチ部分記録発生器と、

上記一次及び二次のスピーチ部分記録発生器で形成された両方のスピーチ部分記録に文法ルールを適用する文法ルール適用サブシステムと、を備えたことを特徴とする装置。

【請求項１０】　ディスプレイ装置と、

入力テキストセグメントが構文的に正しくないという指示を上記ディスプレイ装置に表示させると共に、入力テキストセグメントの識別されたワードが、その識別され

たワードと一般に混同するワードに置き換えられた場合には、入力テキストセグメントが構文的に正しいという指示を上記ディスプレイ装置に表示させるためのフィードバックサブシステムとを更に備えた請求項９に記載の装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、一般に、自然言語パージングの分野に係り、より詳細には、自然言語テキストに生じるエラーを修正する技術に係る。

【０００２】

【従来の技術】文章を書くときに、筆者は、時々、あるワードが正しいところに誤って別のワードを使用することがある。例えば、筆者は、「ａｄ」が正しいところにワード「ａｄｄ」を用いて次のような文章を書くことがある。Ｔｈｅ　ａｄｄ　ｃｏｎｖｉｎｃｅｄ　ｐｅｏｐｌｅ．互いに一貫して間違いを侵す「ａｄｄ」及び「ａｄ」のようなワード対は、一般に混同すると言える。一般に混同するワードは、同様の発音を有する（例えば、「ａｄｖｉｓｅ」対「ａｄｖｉｃｅ」）か、又は若干の文字の位置が異なる（例えば、「ｆｒｏｍ」対「ｆｏｒｍ」）ことが多い。上記例において、ワード「ａｄ」は、「意図されたワード」即ち筆者により意図されたワードと称し、一方、ワード「ａｄｄ」は、「混同するワード」即ち筆者が意図されたワードと誤って置き換えたワードと称する。

【０００３】混同するワードがセンテンスに含まれたときは、自然言語パーザは、センテンスをパージングすることが困難である。自然言語パーザは、自然言語のセンテンスを分析して、センテンスの語彙及び構文内容を見分ける。例えば、チャートに基づく自然言語パーザは、入力センテンスの各ワードに対し辞書から辞書エントリーを検索する。辞書エントリーは、ワードに関する一般的情報を含む語彙記録と、ワードが表すスピーチの特定部分に特有の情報を各々含む参照用のスピーチ部分記録とを備えている。パーザは、１つ以上のスピーチ部分記録を、チャートと称する作用領域に入れ、それらはパージングルールを受け、スピーチ部分記録が、より大きな構文単位、最終的にはセンテンスに合成される。

【０００４】

【発明が解決しようとする課題】自然言語パーザを使用し、意図されたワードとして筆者により意図されたスピーチの部分を有していない混同するワードを含むセンテンスをパージングするときには、自然言語パーザは、センテンスの完全なパーズを発生することができない。自然言語パーザの目的は、入力センテンスの意図された語彙及び構文内容を正確に表す完全なパーズを発生することであるから、混同するワードを含むセンテンスの完全なパーズを発生することのできる自然言語パーザが望まれる。

【０００５】

【課題を解決するための手段】本発明は、センテンスのような入力テキストセグメントのパージング中に一般に混同するワードを識別しそして分析する自然言語パーザを提供する。一般に混同するワードを識別しそして分析する能力は、筆者により作成された文書に含まれたセンテンスの文法上の的確さを評価するためにワードプロセッサに関連して使用される文法チェッカーについて特に価値がある。しかしながら、当業者であれば、本発明は、自然言語パーザのいかなる用途にも有利に使用できることが明らかであろう。

【０００６】本発明は、一般に混同するワードのセットのリストを使用する。各セットは、一般に混同する２つ以上のワードを含んでいる。本発明によるパーザは、入力センテンスにおいて遭遇しこれらセットの１つに現れるワードであって、あたかもそのセットの他のワードで表されるスピーチの部分を表し得るようなワードを処理する。例えば、ワード「ａｄｄ」及び「ａｄ」が混同し得るワードのセットを構成しそしてワード「ａｄｄ」が入力センテンスに現れる場合に、パーザは、ワード「ａｄｄ」をあたかもそれが動詞又は名詞を表し得るかのように処理する。というのは、ワード「ａｄｄ」は動詞を表し、そしてワード「ａｄ」は名詞を表すからである。これは、遭遇するワードの辞書エントリーにスピーチの対応部分がないか、又はスピーチの同じ部分が異なる数又は時制を有するようなセットの他のワードの辞書エントリーからスピーチ部分記録をチャートに追加することを含む。これら付加的なスピーチ部分記録は、一般に、パーザが元のスピーチ部分記録にルールを適用する機会を有した後に、パージングプロセスに後で追加される。

【０００７】本発明の実施形態は、更に、一般に混同するワードを識別及び分析した結果を表示するために文法チェッカーのユーザインターフェイスを提供する。又、ある実施形態において、本発明は、これらの追加されたスピーチ部分記録へのレファレンスを、遭遇したワードの語彙記録に追加し、ワードの語彙記録を用いてワードがスピーチのどの部分を表すかを決定するためのルールが、遭遇したワードがスピーチのこれら追加部分を表す確率を考慮するようにする。

【０００８】

【発明の実施の形態】自然言語パーザにおいて一般に混同するワードを識別しそして分析する方法及びシステムが提供される。好ましい実施形態において、本発明は、一般に混同するワードのセットのリストを使用し、これは、ユーザにより修正することができる。本発明によれば、パーザは、入力センテンスにおいて遭遇しこれらセットの１つに現れるワードであって、あたかもそのセットの他のワードで表されるスピーチの部分を表し得るようなワードを処理する。例えば、ワード「ａｄｄ」及び「ａｄ」が混同し得るワードのセットを構成しそしてワ

ード「ａｄｄ」が入力センテンスに現れる場合に、パーザは、ワード「ａｄｄ」をあたかもそれが動詞又は名詞を表し得るかのように処理する。というのは、ワード「ａｄｄ」は動詞を表し、そしてワード「ａｄ」は名詞を表すからである。これは、遭遇するワードにスピーチの対応部分がないセットの他のワードの辞書エントリーからスピーチ部分記録をチャートに追加することを含む。これらの付加的なスピーチ部分記録は、一般に、パーザが元のスピーチ部分記録にルールを適用する機会を有した後に、パージングプロセスに後で追加される。

又、ある実施形態では、本発明は、これらの追加されたスピーチ部分記録へのレファレンスを、遭遇したワードの語彙記録に追加し、従って、ワードの語彙記録を用いてワードがスピーチの他のどの部分を表すかを決定するためのルールが、遭遇したワードがスピーチのこれら追加部分を表す確率を考慮するようにする。

【０００９】図１は、パーザが好ましく動作する汎用コンピュータシステムの高レベルブロック図である。コンピュータシステム１００は、中央処理ユニット（ＣＰＵ）１１０と、入力／出力装置１２０と、コンピュータメモリ（メモリ）１３０とを備えている。入力／出力装置の中には、ハードディスクドライブのような記憶装置１２１がある。又、入力／出力装置は、取り外し可能なメディアドライブ１２２を含み、このドライブは、適用パーザを含むソフトウェア製品をインストールするのに使用でき、これらソフトウェア製品は、ＣＤ－ＲＯＭのようなコンピュータ読み取り可能な媒体に設けられる。更に、入力／出力装置は、ユーザが自然言語テキストを直接的に入力するのに使用するキーボード１２３も含む。又、入力／出力装置は、ユーザが自然言語テキストを間接的に入力するのに使用する音声入力装置１２４及び手書き入力装置１２５も任意に含む。音声入力装置を用いてユーザにより入力される自然言語テキストは、好ましくは、音声認識装置（図示せず）により音声データから変換される。同様に、手書き入力装置を用いてユーザにより入力される自然言語テキストは、好ましくは、手書き認識装置（図示せず）を用いて手書きデータから変換される。メモリ１３０は、一般的に混同するワードを識別及び分析するためのパーザ１３１を含む。パーザは、入力テキストセグメント及び中間パーズ結果を表すパーズツリーを含むためのチャート１３２を備えている。又、パーザは、一般に混同するワードを、それらと一般に混同するワード（即ち、意図されたワード）へとマップする混同し得るワードのテーブル１３３も備えている。例えば、ワード「ａｄｄ」は、ワード「ａｄ」へとマップされ、筆者が「ａｄ」ではなくワード「ａｄｄ」を間違って使用するかもしれないことを指示する。混同し得るワードのテーブルは、記憶装置に記憶されてもよいし、又は取り外し可能なメディアドライブを用いて取り外し可能な媒体に記憶されてもよい。パーザは、上記のように構成されたコンピュータシステムにおいて実施されるのが好ましいが、異なる構成のコンピュータシステムでも実施できることが当業者に明らかであろう。

【００１０】適用パーザを用いて入力テキストをパーズしながら、一般に混同するワードを識別及び分析する一例を、図２ないし７に関連して説明する。図２は、入力テキストに生じる潜在的に混同するワードを含む入力テキストのワードに対しパーザがスピーチ部分記録(part-of-speech record) をチャートに追加するところを示すチャート図である。図２は、例示的入力ストリング２０１「Ｔｈｅ　ａｄｄｃｏｎｖｉｎｃｅｄ　ｐｅｏｐｌｅ．」を示している。図２は、更に、パーザが入力テキストに現れるワードに対するスピーチ部分記録をチャートに追加した後のパーザのチャート２００の内容も示している。このチャートは、スピーチ部分記録２１１－２１５を含み、その１つ以上は、入力テキストに生じる各ワードを示す。スピーチ部分記録２１１は、ワード「ｔｈｅ」を表し、スピーチ部分記録２１２は、ワード「ａｄｄ」を表し、スピーチ部分記録２１３は、ワード「ｃｏｎｖｉｎｃｅｄ」を表し、そしてスピーチ部分記録２１４及び２１５は、ワード「ｐｅｏｐｌｅ．」を表す。各スピーチ部分記録は、そのワードが表し得るスピーチの１つの考えられる部分の指示と、動詞のスピーチ部分記録に対する動詞の時制のような付加的な関連語彙情報とを含む。スピーチ及び他の語彙情報の考えられる部分は、入力ストリングに生じるワードに対し辞書入力から検索されるのが好ましい。

【００１１】図３は、チャートの内容により暗示されるルールをパーザが適用するところを示す。即ち、パーザは、チャートに既に存在する形式の記録を結合することのできるルールを適用する。図３は、動詞及び名詞を動詞句即ち「ＶＰ」に変換するルールの適用を示す。このルールの適用は、ワード「ｃｏｎｖｉｎｃｅｄ」に対する動詞のスピーチ部分記録３１３を「ｐｅｏｐｌｅ」に対する名詞のスピーチ部分記録３１４と結合する動詞句記録３２１を形成する。スピーチ部分記録及びルーツにより成形された記録の両方は、付加的なルールを包含する。これら付加的なルールは、繰り返し適用される。ここに示す例の場合には、チャートの記録を結合するのに他のルールは首尾良く適用されない。

【００１２】これらルールの適用により完全なパーズが発生された場合には（即ち、入力ストリングの全てのワードをカバーするセンテンス記録が形成された場合には）、パージングが終了しそして完全なパーズが返送されるが、さもなくば、パーザの動作が続けられる。入力テキストが潜在的に混同するワードを含む場合には、パーザの動作が続けられるが、さもなくば、パーザは完全なパーズを発生することができず、欠陥を返送する。図４は、入力テキストにおいて潜在的に混同するワードを

識別するのに使用される例示的な混同し得るワードのテーブルを示す図である。この混同し得るワードのテーブル４００は、潜在的に混同するワードの欄と、考えられる意図されたワードの欄とを含む。各行において、潜在的に混同するワードの欄は、１つ以上の他のワードに対して混同することのあるワードを含む。その行において、考えられる意図されたワードの欄は、潜在的に混同するワードと混同し得る１つ以上の考えられる意図されたワードのリストを含む。例えば、行４０２は、ワード「ａｄｄ」がワード「ａｄ」と混同し得ることを示す。行４０１は、その逆もあることを示し、即ちワード「ａｄ」がワード「ａｄｄ」と混同することを示している。ワード間の幾つかの潜在的な混同は、一方向性であり、即ちある対の一方のワードは、その対の他方のワードと潜在的に混同し得る（例えば、行４０３は、「ｃａｎｔ」が「ｃａｎ’ｔ」と混同し得ることを示す）が、その逆はない（例えば、「ｃａｎ’ｔ」は、潜在的に混同するワードの欄に現れない）。行４０９ないし４１１は、潜在的に混同するワードが２つ以上の考えられる意図されたワードと混同し得ることを示す。パーザは、入力テキストのワードを、混同し得るワードのテーブルの潜在的に混同するワード欄のワードと比較する。入力ストリングのいずれかのワードが、潜在的に混同するワード欄のワードと一致する場合には、その入力テキストが潜在的に混同するワードを含む。

【００１３】図５は、入力テキストにおいて識別された潜在的に混同するワードに対応する考えられる意図されたワードに対しパーザがスピーチ部分記録をチャートに追加するところを示すチャート図である。パーザは、潜在的に混同するワード以外のスピーチ部分を有する考えられる意図されたワードに対するスピーチ部分記録をチャートに追加するのが好ましい。例えば、パーザは、好ましくは、考えられる意図されたワード「ａｄ」に対し名詞のスピーチ部分記録を追加する。というのは、そのスピーチ部分が、その潜在的に混同するワード「ａｄｄ」に対して考えられるスピーチ部分とは異なるからである。又、パーザは、好ましくは、潜在的に混同するワードとは異なる時制を有する考えられる意図されたワードに対しスピーチ部分記録をチャートに追加する。例えば、パーザは、好ましくは、考えられる意図されたワード「ｍｉｎｄ」に対し現在時制の動詞のスピーチ部分記録を追加する。というのは、その時制が、潜在的に混同するワード「ｍｉｎｅｄ」の過去時制の動詞形態と異なるからである。又、パーザは、好ましくは、潜在的に混同するワードとは異なる数を有する考えられる意図されたワードに対しスピーチ部分記録をチャートに追加する。例えば、パーザは、好ましくは、考えられる意図されたワード「ｌａｐｓ」に対し複数名詞のスピーチ部分記録を追加する。というのは、その数が、潜在的に混同するワード「ｌａｐｓｅ」の単数名詞形と異なるからである。更に、混同するワードの特定セットに対し、ユーザは、たとえスピーチ部分、時制及び数が同じであっても、セット内のワードに対しスピーチ部分記録をチャートに追加することを指定するのが好ましい。図５は、パーザがワード「ａｄ」に対する名詞の市ピーチ部分記録５１６をチャートに追加したところを示す。というのは、行４０２に示すように、入力ストリングに現れるワード「ａｄｄ」が、辞書が考えられるスピーチ部分として名詞を指定するところのワード「ａｄ」と混同し得るからである。

【００１４】図６は、考えられる意図されたワードに対してスピーチ部分記録をチャートに追加した後にチャートの内容により暗示されるルールをパーザが適用するところを示すチャート図である。図６は、チャートに記録６２２及び６２３を形成するルールをパーザが適用するところを示す。記録６２２は、「ｔｈｅ」に対する冠詞のスピーチ部分記録６１１を、考えられる意図されたワード「ａｄ」に対する名詞のスピーチ部分記録６１６と結合して、名詞句（ＮＰ）を形成する。記録６２３は、名詞句の記録６２２と動詞句の記録６２１を結合してセンテンスを形成する。又、記録６２３は、入力テキストの各ワードを表すリーフを有するツリーのヘッドノードを構成するという点で、入力ストリングの各ワードを「カバー」する。

【００１５】別のワードに対するスピーチ部分記録を追加した後のチャートの内容によって暗示されたルールの適用が完全なパーズを形成した場合には、パーザは、的確な成功を返送するが、さもなくば、パーザは失敗を返送する。図６から明らかなように、この例では、パーザは、センテンス記録６２３が入力テキストの全てのワードをカバーし、それ故、的確な成功を返送するという点で、完全なパーズを形成している。

【００１６】図７は、パーザを用いた文法チェッカーの視覚的ユーザインターフェイスを示すスクリーン図である。この文法チェッカーのユーザインターフェイスは、好ましくは、ウインドウ７００を表示する。ウインドウ７００は、好ましくは、現在チェックされているセンテンス７１０を含む。更に、ウインドウは、現在センテンスの特定のワードがおそらく別のワードと混同しているという指示７２０を含む。又、ウインドウは、好ましくは、潜在的に混同するワードを置き換えるという提案７１０も含む。更に、ウインドウは、好ましくは、提案を受け入れて潜在的に混同するワードを置き換えることをユーザが選択できるボタン７４０と、提案を拒否しそして潜在的に混同するワードの置き換えを排除することをユーザが選択できるボタン７５０とを含む。

【００１７】図８は、一般に混同するワードを識別しそして分析しながら入力テキストをパーズするために適用パーザにより好ましく実行される高レベルステップを示すフローチャートである。ステップ８０１において、パ

ーザは、入力テキストに生じる潜在的に混同するワードを含む入力テキストのワードに対しスピーチ部分記録をチャートに追加する。ステップ８０２において、パーザは、チャートの内容により暗示されたルールの１つを適用する。ステップ８０３において、ステップ８０２のルールの適用により完全なパーズが形成された場合には（即ち、入力ストリングの全てのワードをカバーするセンテンス記録が形成された場合には）、これらのステップは終了し、完全なパーズが返送されるが、さもなくば、パーザはステップ８０４に続く。ステップ８０４において、パージングが終了した場合、即ち暗示された全てのルールが適用されるか又は適用されたルールの全数が上限を越えた場合には、パーザは、ステップ８０５に続き、さもなくば、パーザは、ステップ８０２に続いて、暗示された別のルールを適用する。ステップ８０５において、入力テキストが潜在的に混同するワードを含む場合には、パーザは、ステップ８０６に続き、さもなくば、パーザは、完全なパーズを形成することができず、失敗を返送する。入力テキストが潜在的に混同するワードを含むかどうか決定するために、パーザは、入力テキストのワードを、混同し得るワードのテーブルの潜在的に混同するワード欄のワードと比較する。入力ストリングのいずれかのワードが、潜在的に混同するワード欄のワードと一致する場合には、入力テキストが潜在的に混同するワードを含む。ステップ８０６において、パーザは、入力テキストにおいて識別された潜在的に混同するワードに対応する考えられる意図されたワードに対し、スピーチ部分記録をチャートに追加する。ステップ８０７において、パーザは、別のワードに対しスピーチ部分記録をステップ８０６で追加した後にチャートの内容により暗示されたルールの１つを適用する。ステップ８０８において、完全なパーズが発生された場合には、パーザは、的確な成功を返送し、さもなくば、パーザは、ステップ８０９に続く。ステップ８０９において、パージングが完了した場合、即ち暗示された全てのルールが適用されるか又は適用されたルールの全数が上限を越えたな場合に、パーザは、失敗を返送し、さもなくば、パーザはステップ８０７に続き、別の暗示されたルールを適用する。

【００１８】幾つかの文法ルールは、各ワードに対してスピーチの特定部分に適用される間に、各ワードが表すスピーチの全ての考えられる部分を考慮する。このような文法ルールは、ルールを適用するのに必要であるが完全なパーズの発生に貢献しそうもない処理リソースの量を減少することができる。このようなリソースをサポートするために、スピーチ部分記録を一緒にリンクし、ワードに対して考えられる全てのスピーチ部分を容易に決定することができる。本発明の好ましい実施形態によれば、考えられる意図されたワードに対するスピーチ部分記録は、パージングプロセスの始めに、潜在的に混同するワードに対するスピーチ部分記録にリンクされるのが好ましい。図８ないし１１は本発明のこの特徴を示す。図９は、図２の別の図であって、ステップ８０１の実行の後であって且つルールを適用する前のチャートの内容を示す。入力テキストのワードにより表されたスピーチの潜在的な部分に関するデータを含むのではなく、スピーチ部分記録９１１ないし９１５は、この情報を含むデータ構造体へのポインタを含む。例えば、スピーチ部分記録９１４は、潜在的なスピーチ部分の名詞及び他の関連する語彙情報を含むスピーチ部分データ構造体９７１へのポインタを含む。別の好ましい実施形態（図示せず）によれば、スピーチ部分データ構造体へのポインタをデレファレンスする時間的コストを排除するために、スピーチ部分データ構造体からスピーチ部分記録へデータがコピーされる。スピーチ部分データ構造体９７１は、ワード「ｐｅｏｐｌｅ」を表す語彙記録９７０への両方向性リンクを含む。別のデータ構造体９７２は、ワード「ｐｅｏｐｌｅ」に対して考えられるスピーチ部分の動詞を含むと共に、語彙記録９７０への両方向性リンクも含む。スピーチ部分データ構造体９７１及び９７２と語彙記録データ構造体９７０との間のリンクは、ルールがそれらの処理を、特定のワードに対して考えられる全てのスピーチ部分のセットに基づいて行えるようにする。ワード「ｐｅｏｐｌｅ」に対する名詞のスピーチ部分記録９１４の場合に、この記録に適用されるルールは、ワード「ｐｅｏｐｌｅ」が動詞も表し得ることを考慮する。

【００１９】図１０は、本発明のこの特徴によるチャートの更に別の変形を示している。図１０は、考えられる意図されたワードに対するスピーチ部分データ構造体と、その潜在的に混同するワードに対する語彙記録データ構造体とのリンクを示す。図１０は、名詞形態のワード「ａｄ」に対するスピーチ部分データ構造体１０５２を「ａｄｄ」に対する語彙記録１０５０に加えるところを示している。両スピーチ部分データ構造体１０５１及び１０５２は、ワード「ａｄｄ」の語彙データ構造体１０５０に両方向性リンクされるので、「ａｄｄ」に対する動詞のスピーチ部分記録１０１２に適用されるルールは、このワードに対して考えられる名詞のスピーチ部分とみなすことができる。上記のように、本発明によれば、考えられる意図されたワードの考えられるスピーチ部分に対するスピーチ部分データ構造体は、ルールにより結合されるべきワードに対して考えられる他のスピーチ部分に基づいて作用するルールの適用を容易にするために、各々の考えられる混同するワードに対する語彙データ構造体にリンクされるのが好ましい。図１１は、本発明のこの特徴により、名詞のスピーチ部分記録１１１６がステップ８０６においてチャートに追加されたときに、考えられる意図されたワード「ａｄ」に対する潜在的なスピーチ部分の名詞を含むスピーチ部分データ構造

体１１５２へのポインタを含むことを示している。

【００２０】完全なパーズの発生に貢献しそうにないルールの適用を防止することによってパージングの効率を高めるのに加えて、考えられる意図されたワードに対してリンクされるスピーチ部分記録を参照するルールを使用することは、潜在的に混同するワードが考えられる意図されるワードと実際に混合する場合にパーザが潜在的に混同するワードを用いて入力テキストの見掛け上正しい完全なパーズを発生するのを防止することができる。これは、混同するワード「ｆｏｒｍ」が意図されたワード「ｆｒｏｍ」に代わって使用される次の例示的なセンテンスについて言えることである。Ａｎｇｅｌａ ｄｅｐａｒｔｅｄ ｆｏｒｍ Ｓｅａｔｔｌｅ．考えられる意図されたワードの考えられるスピーチ部分を考慮しないルールを用いると、あるパーザは、動詞句が動詞句「ｄｅｐａｒｔｅｄ」と名詞句「ｆｏｒｍ Ｓｅａｔｔｌｅ」から形成されるこのセンテンスの完了パーズを発生する。この完了パーズは、「ｆｏｒｍ Ｓｅａｔｔｌｅ」が動詞「ｄｅｐａｒｔｅｄ」の有効な目的語でないから、実際には正しくない。しかしながら、考えられる意図されたワードの考えられるスピーチ部分を考慮するルールを用いると、パーザは、この誤った完了パーズを回避することができる。この場合に、動詞句と動詞句の目的語である名詞句を結合して別の動詞句にするルールを適用することは、名詞句の「前修飾語」（即ち、名詞句のメインワード「Ｓｅａｔｔｌｅ」の前に生じるワード「ｆｏｒｍ」）又はその考えられる意図されたワードがスピーチの前置詞部分を表し得るときに、スピーチの前置詞部分が動詞の目的語の前に意図される可能性が大きい場合は、阻止される。考えられる意図されたワード「ｆｒｏｍ」は、スピーチの前置詞部分をもつことができ、そして前置詞のスピーチ部分記録は、潜在的に混同するワード「ｆｏｒｍ」のスピーチ部分記録にリンクされるので、このルールの適用は阻止され、パーザが上記の誤った完了パーズを形成しないよう防止し、これにより、パーザの出力の精度を改善する。

【００２１】図１２及び１３は、潜在的に混同するワードのスピーチ部分記録がチャートに追加されそして暗示されたルールがステップ８０５によりそれらに適用された後に、別のワードに対するスピーチ部分記録をチャートに追加させる２つの好ましい実施形態を示している。図１２は、２つの異なるリスト即ち「待ち行列」である一次リスト１２８０及び二次リスト１２９０からスピーチ部分記録がチャート１２０に追加される実施形態を示す。入力テキストに含まれたワードに対するスピーチ部分記録は、一次リスト１２８０に記憶される。この一次リストは、入力テキストに現れるワード「ｔｈｅ」、「ａｄｄ」、「ｃｏｎｖｉｎｃｅｄ」及び「ｐｅｏｐｌｅ」に対するスピーチ部分記録を含む。二次リスト１２９０は、考えられる意図されたワードに対するスピーチ部分記録を含む。二次リスト１２９０は、別のワード「ａｄ」に対する名詞のスピーチ部分記録を含むことが明らかである。この実施形態では、ワードは、先ず、一次リストからチャートに追加される。暗示されたルールが適用された後に、パーザは、二次リストからスピーチ部分記録をチャートに追加する。好ましい実施形態において、スピーチ部分記録を二次リストからチャートに追加することは、先ず、二次リストから一次リストへそれらを移動し、次いで、それらを一次リストからチャートへ追加する一方、一次リストに現れる新たに暗示されるルールを適用することを含む。この解決策は、入力テキストのワードと一般に混同するところの考えられる意図されたワードがチャートに追加される前に、入力テキストに含まれたワードに対するスピーチ部分記録からパーズツリーを構成できるようにする。

【００２２】図１３は、スピーチ部分記録が単一のリスト１３７０からチャート１３００に追加される別の実施形態を示す。リスト１３７０は確率リストであり、完了パーズツリーのリーフを最終的に構成する各スピーチ部分記録の確率に基づいて分類される。確率指向パーザの詳細な説明については、参考としてここに取り上げる「統計学的な処理をルールに基づく自然言語パーザへとブートストラップするための方法及びシステム(METHOD AND SYSTEM FOR BOOTSTRAPPING STATISTICAL PROCESSING INTO A RULE-BASED NATURAL LANGUAGE PARSER)」と題する米国特許出願第０８／２６５，８４５号を参照されたい。スピーチ部分記録は、成功裡なパーズツリーのリーフを構成する確率の下降順に確率リストからチャートに追加される。これらの確率は、「適用優先値」とも称されるが、入力テキストセグメントの代表的集成に対して完成したパーズツリーにおける各スピーチ部分記録の出現を統計学的に分析することによって発生されるのが好ましい。例えば、スピーチ部分記録１３７４及び１３７５に関連して示された統計データは、ワード「ｐｅｏｐｌｅ」を含む入力テキストセグメントにおいて、ワード「ｐｅｏｐｌｅ」が入力セグメントの完全なパーズにおいて名詞を表すときが７８％であり、一方、動詞を表すケースが１３％であることを示している。この実施形態において、考えられる意図されたワードに対するスピーチ部分記録には、比較的小さな確率が指定され、それらを処理の終わり付近でチャートに追加させるのが好ましい。これは、多数の方法で行われる。ワード「ａｄ」は入力テキストに実際に生じないので、入力テキストに「ａｄ」が現れるときにワード「ａｄ」の名詞形態が完全なパーズツリーのリーフを形成する確率は、低くすることができる。或いは又、ワード「ａｄ」ではなくワード「ａｄｄ」が入力テキストに現れるときにワード「ａｄ」の名詞形態が完成パーズツリーのリーフを形成する確率を計算するように個別の統計データを維持することもできる。いずれにせよ、代替えワード「ａｄ」に対す

るスピーチ部分記録は、潜在的に混同するワード「ａｄｄ」に対するスピーチ部分記録の後にチャートに追加される。最後に、考えられる意図されたワードに対するスピーチ部分記録の確率は、確率リストの最少確率以下にセットすることができる。

【００２３】更に別の好ましい実施形態では、潜在的に混同するワードに対するスピーチ部分記録がチャートに追加された後に代替えワードに対するスピーチ部分記録をチャートに追加させる２つの解決策を組み合わせて、潜在的に意図されたワードに対するスピーチ部分記録を二次リストに記憶し、そして全てのルールと、一次リストで終わりとなるスピーチ部分記録とをそれらの確率により順序付けする。

【００２４】ユーザは、潜在的に混同するワード、又はその潜在的に混同するワードに置き換わる考えられる意図されたワードのリストを変更するようにパーザを構成できるのが好ましい。ユーザは、図４に示された混同し得るワードのテーブルを変更することによりこれを行う。しかしながら、あるユーザは、混同し得るワードのリストの簡単な表示を変更できることを望む。図１４は、簡単化された混同し得るワードのファイル１４００を示す。行１４０１ないし１４０６は、潜在的に混同するワードの１つのセットに各々対応する。特に指示のない限り、１つの行に一緒に現れるワードは、全て、互いに混同する。例えば、行１４０１は、ワード「ａｄ」がワード「ａｄｄ」と混同しそしてその逆もあり得ることを示す。又、この混同し得るワードのファイルは、潜在的に混同するとみなしてはならないワードの前にハイフォン（－）記号を置くことによりユーザが一方向の混同関係を指定できるのが好ましい。例えば、行１４０１におけるワード「ｃａｎ’ｔ」の前のハイフォン記号は、ワード「ｃａｎｔ」がワード「ｃａｎ’ｔ」と混同することはあるが、ワード「ｃａｎ’ｔ」がワード「ｃａｎｔ」と潜在的に混同されないことを示す。又、この混同し得るワードのファイルは、潜在的に混同するセットのワードは、それらが同じスピーチ部分、数及び時制を有していても、互いに置き換えられることをユーザがアスタリスク（＊）記号で指定できるのが好ましい。例えば、混同し得るセット１４０２の前のアスタリスク記号は、「ｃａｎ’ｔ」及び「ｃａｎｔ」の両方のワードがスピーチの動詞部分を有していても、ワード「ｃａｎ’ｔ」をワード「ｃａｎｔ」に置き換えられることを示す。パーザは、ユーザがパーザの動作を構成するように混同し得るワードのファイル１４００を変更できると共に、混同し得るワードのファイルを、図４に示す混同し得るワードのテーブルのようなパーザにより容易に適用される形態へと変換できるのが好ましい。

【００２５】好ましい実施形態を参照して本発明を説明したが、当業者であれば、本発明の範囲から逸脱せずに、種々の変更や修正がなされ得ることが明らかであろう。例えば、上記以外の機構を使用して、一般に混同するワードに対するスピーチ部分記録をチャートに導入することができる。更に、上記した本発明の実施形態は、コンピュータプログラミング言語やテキスト形成言語のような人為的言語のテキストをパーズするのにも容易に適用できる。

【図面の簡単な説明】

【図１】本発明が好ましく動作する汎用コンピュータシステムの高レベルブロック図である。

【図２】入力テキストに生じるワードに対してパーザがスピーチ部分記録をチャートに追加するところを示すチャート図である。

【図３】チャートの内容により暗示されるルールをパーザが適用するところ示すチャート図である。

【図４】入力テキストにおける潜在的に混同するワードを識別するのに使用される例示的な混同ワードテーブルを示すテーブル図である。

【図５】入力テキストにおいて識別された潜在的に混同するワードに対応する考えられる意図されたワードに対し、パーザがスピーチ部分記録をチャートに追加するところを示すチャート図である。

【図６】考えられる意図されたワードに対するスピーチ記録の部分がチャートに追加された後に、チャートの内容により暗示されたルールをパーザが適用するところを示すチャート図である。

【図７】パーザを用いた文法チェッカーの視覚的ユーザインターフェイスを示すスクリーン図である。

【図８】一般に混同するワードを識別しそして分析しながら入力テキストをパーズするために適用パーザにより好ましく実行される高レベルステップを示すフローチャートである。

【図９】辞書からの語彙記録を含む図２の別の図であって、ステップ８０１の実行後で且つルールを適用する前のチャートの内容を示す図である。

【図１０】代替えワードに対するスピーチ部分データ構造体を、その潜在的に混同するワードに対する語彙記録データ構造体にリンクするところを示す図である。

【図１１】考えられる意図されたワードに対するスピーチ部分記録を追加した後のチャートの内容を示す図である。

【図１２】スピーチ部分記録が２つの異なるリストからチャートに追加される実施形態を示す図である。

【図１３】スピーチ部分記録が、確率でランク付けされた単一リストからチャートに追加される別の実施形態を示す図である。

【図１４】混同し得るワードの簡単化されたファイルを示す図である。

【符号の説明】

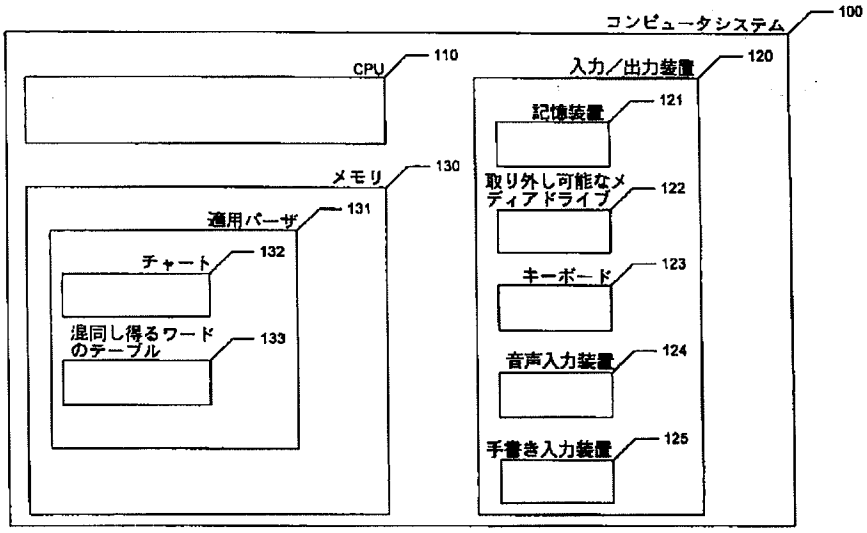
１００　コンピュータシステム

１１０　中央処理ユニット

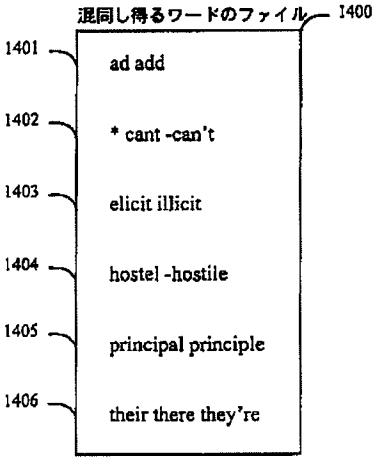
15　　16

１２０　入力／出力装置
１２１　記憶装置
１２２　取り外し可能なメディアドライブ
１２３　キーボード
１２４　音声入力装置
１２５　手書き入力装置
１３０　コンピュータメモリ
＊１３１　パーザ
１３２　チャート
１３３　混同し得るワードのテーブル
２００　パーザのチャート
２０１　入力ストリング
２１１－２１５　スピーチ部分記録
＊

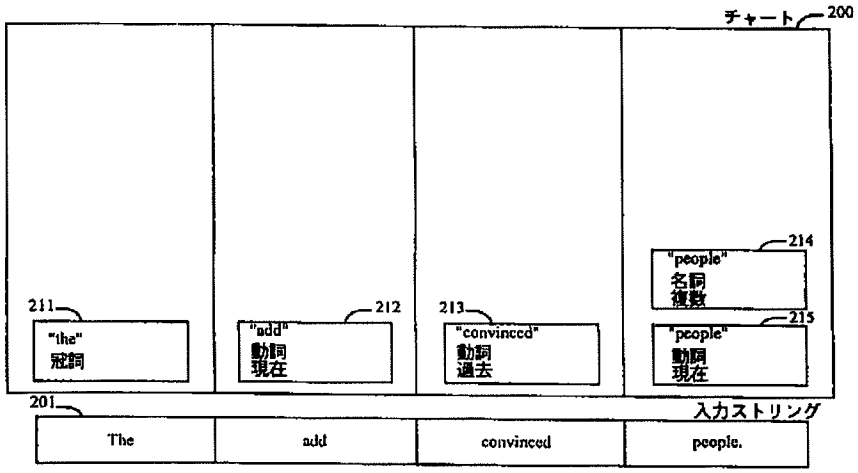
【図１】



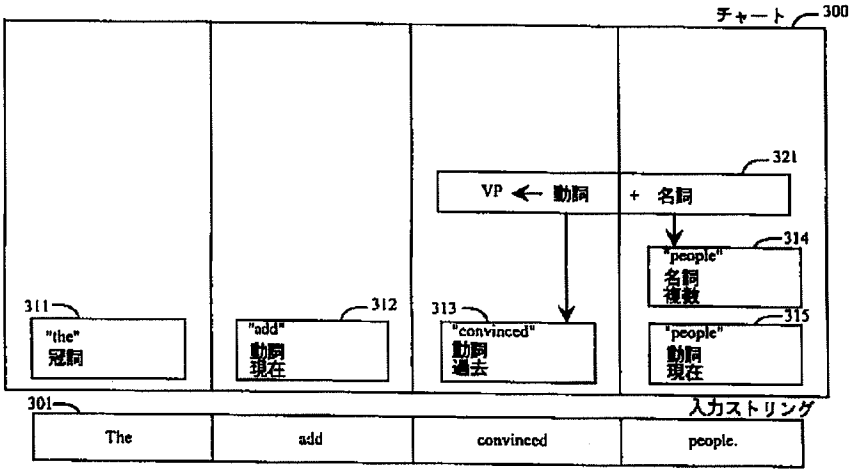
【図１４】



【図２】



【図３】



【図４】

混同し得るワードのテーブル 400

潜在的に混同するワード	考えられる意図されたワード	
ad	add	401
add	ad	402
cant	can't	403
elicit	illicit	404
hostel	hostile	405
illicit	elicit	406
principal	principle	407
principle	principal	408
their	there, they're	409
there	their, they're	410
they're	their, there	411

【図５】

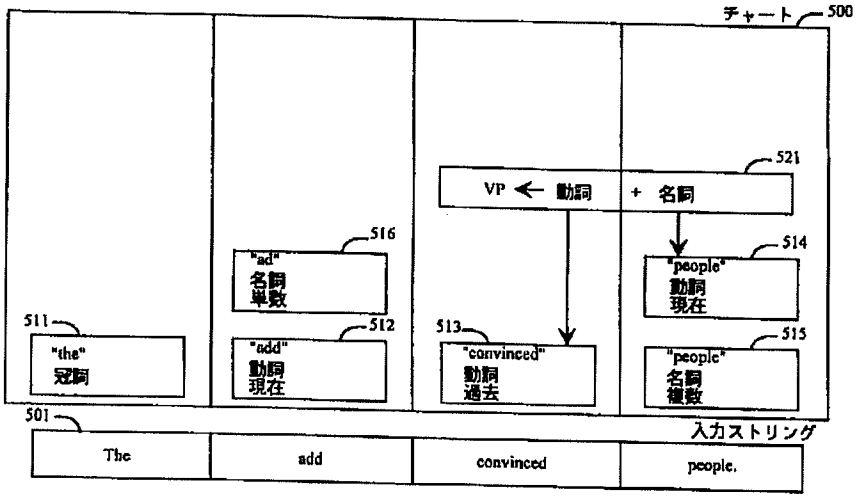


チャート
500
521
VP ← 動詞 + 名詞
516
"ad"
名詞
単数
514
"people"
動詞
現在
511
"the"
冠詞
512
"add"
動詞
現在
513
"convinced"
動詞
過去
515
"people"
名詞
複数
501
入力ストリング
The
add
convinced
people.

【図６】

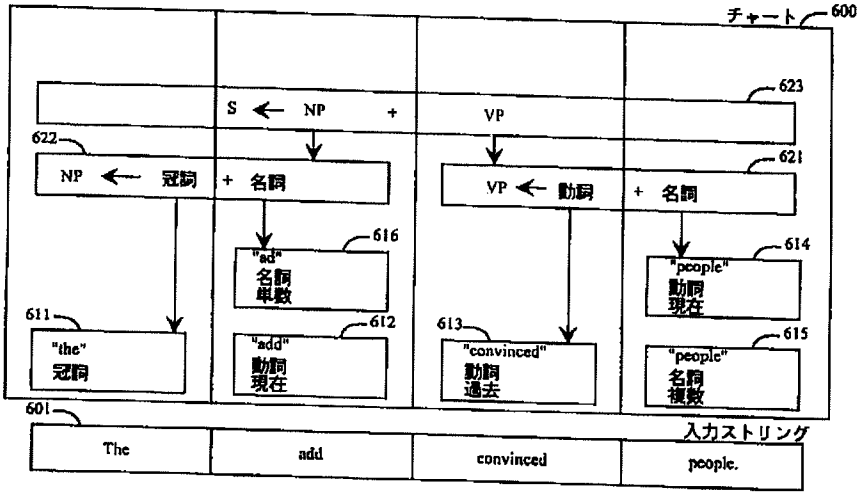
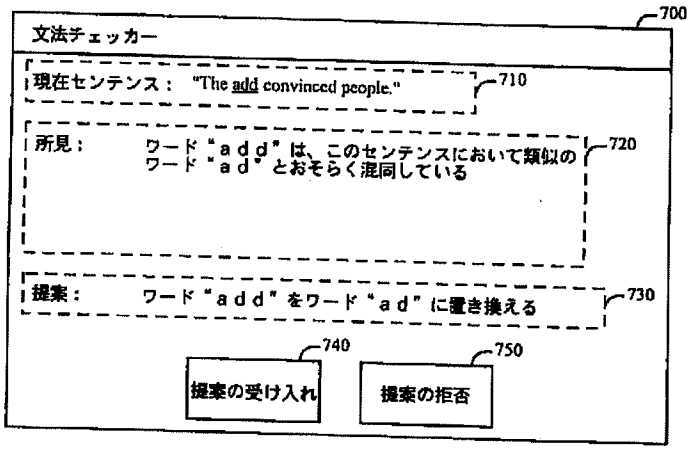
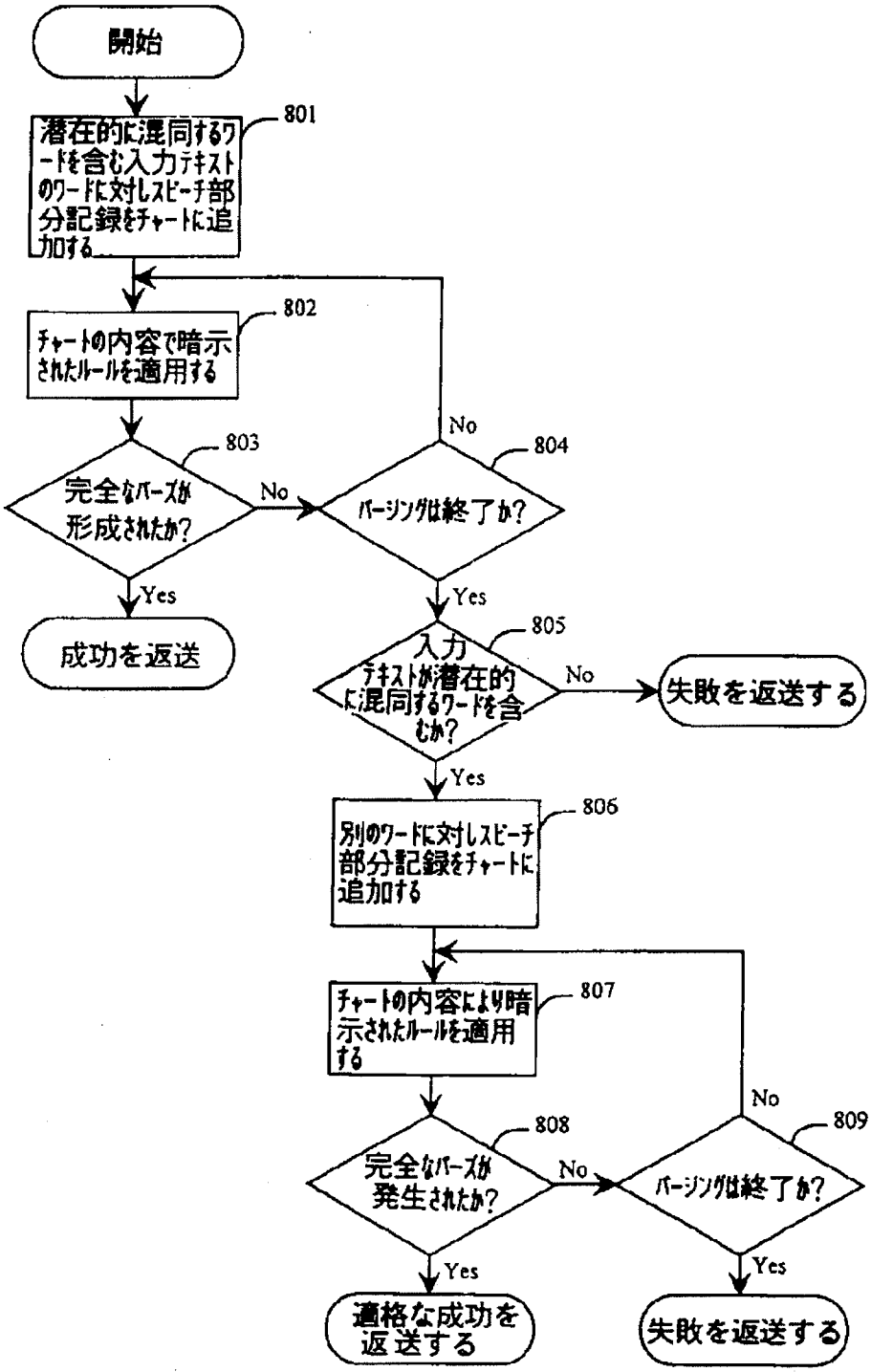


チャート
600
623
S ← NP + VP
622
NP ← 冠詞 + 名詞
621
VP ← 動詞 + 名詞
616
"ad"
名詞
単数
614
"people"
動詞
現在
611
"the"
冠詞
612
"add"
動詞
現在
613
"convinced"
動詞
過去
615
"people"
名詞
複数
601
入力ストリング
The
add
convinced
people.

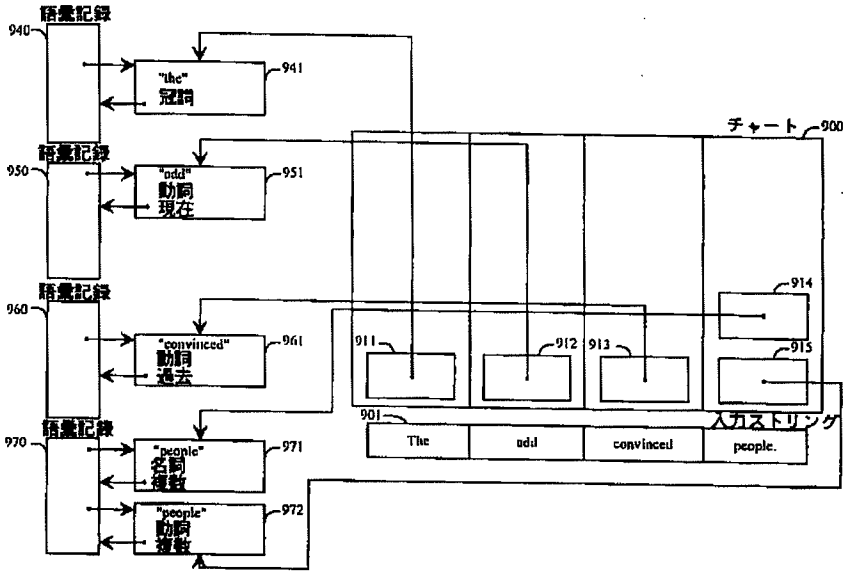
【図７】


700
文法チェッカー
現在センテンス： "The add convinced people."
710
所見： ワード "ａｄｄ" は、このセンテンスにおいて類似の
ワード "ａｄ" とおそらく混同している
720
提案： ワード "ａｄｄ" をワード "ａｄ" に置き換える
730
740
提案の受け入れ
750
提案の拒否

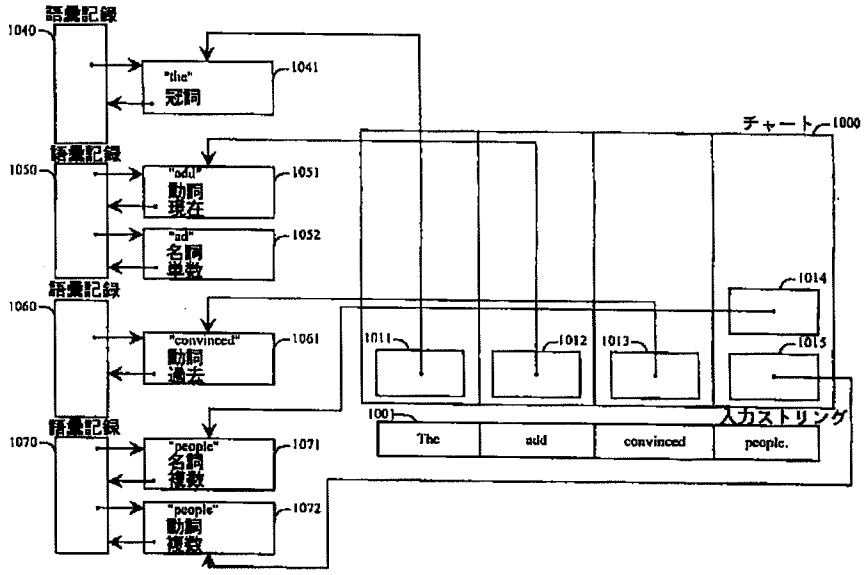
【図８】


開始
801
潜在的に混同するワードを含む入力テキストのワードに対しスピーチ部分記録をチャートに追加する
802
チャートの内容で暗示されたルールを適用する
803
完全なパーズが形成されたか?
No
804
パージングは終了か?
No
Yes
Yes
成功を返送
805
入力テキストが潜在的に混同するワードを含むか?
No
失敗を返送する
Yes
806
別のワードに対しスピーチ部分記録をチャートに追加する
807
チャートの内容により暗示されたルールを適用する
808
完全なパーズが発生されたか?
No
809
パージングは終了か?
No
Yes
Yes
適格な成功を返送する
失敗を返送する

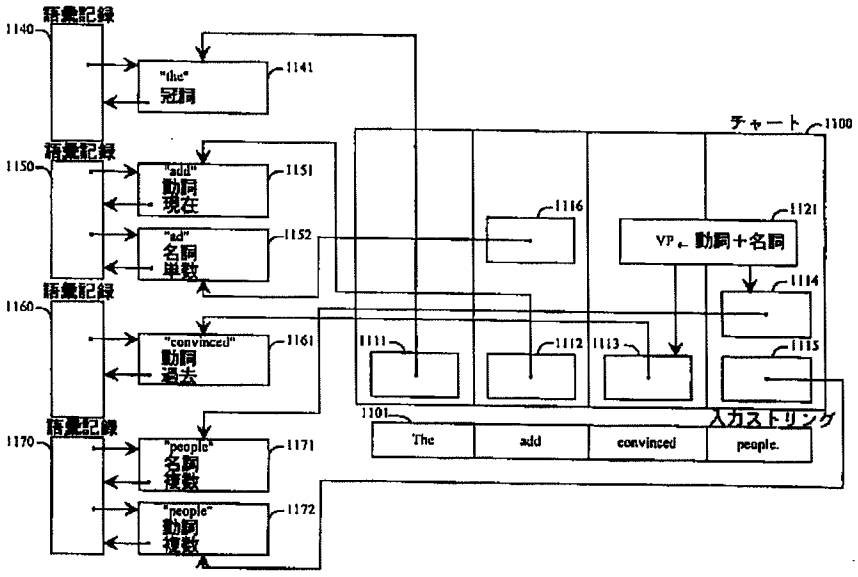
【図９】


語彙記録
940
"the"
冠詞
941
語彙記録
950
"add"
動詞
現在
951
語彙記録
960
"convinced"
動詞
過去
961
語彙記録
970
"people"
名詞
複数
971
"people"
動詞
複数
972
チャート
900
911
912
913
914
915
901
The
add
convinced
people.
入力ストリング

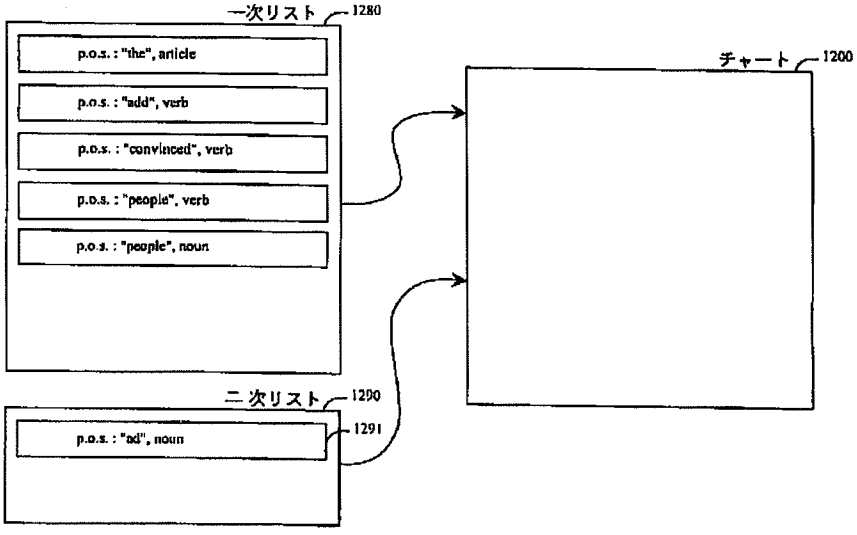
【図１０】


語彙記録
1040
"the"
冠詞
1041
語彙記録
1050
"add"
動詞
現在
1051
"ad"
名詞
単数
1052
語彙記録
1060
"convinced"
動詞
過去
1061
語彙記録
1070
"people"
名詞
複数
1071
"people"
動詞
複数
1072
チャート
1000
1011
1012
1013
1014
1015
100
The
add
convinced
people.
入力ストリング

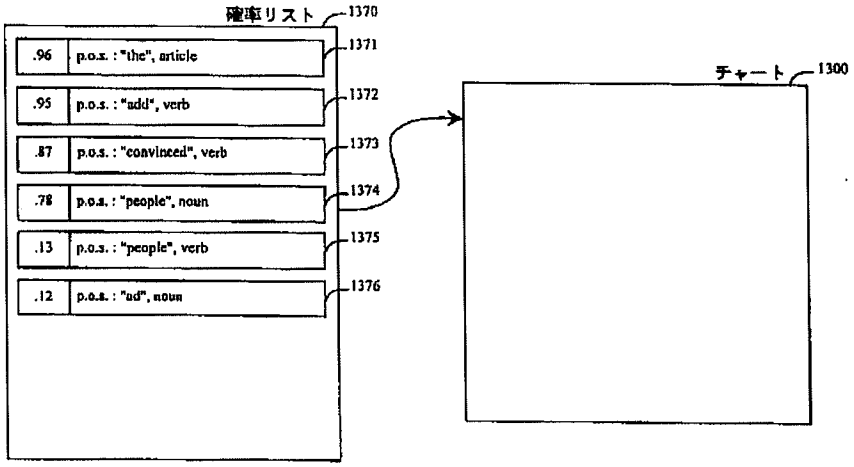
【図１１】


語彙記録
1140
"the"
冠詞
1141
語彙記録
1150
"add"
動詞
現在
1151
"ad"
名詞
単数
1152
語彙記録
1160
"convinced"
動詞
過去
1161
語彙記録
1170
"people"
名詞
複数
1171
"people"
動詞
複数
1172
チャート
1100
1116
1121
VP ← 動詞＋名詞
1114
1111
1112
1113
1115
1101
The
add
convinced
people.
入力ストリング

【図１２】


一次リスト
1280
p.o.s. : "the", article
p.o.s. : "add", verb
p.o.s. : "convinced", verb
p.o.s. : "people", verb
p.o.s. : "people", noun
チャート
1200
二次リスト
1290
p.o.s. : "ad", noun
1291

【図１３】



フロントページの続き

(72)発明者　ジョージ　イー　ヘイドーン
アメリカ合衆国　ワシントン州　98008
ベルビュー　ワンハンドレッドアンドシッ
クスティフィフス　プレイス　ノースイー
スト　3211